Detection of Functionally Important Regions in “Hypothetical Proteins” of Known Structure

Guy Nimrod,[1] Maya Schushan,[1] David M. Steinberg,[2] and Nir Ben-Tal[1,*]
[1]Department of Biochemistry, George S. Wise Faculty of Life Sciences
[2]Department of Statistics and Operations Research, Raymond and Beverly Sackler Faculty of Exact Sciences
Tel Aviv University, 69978 Tel Aviv, Israel
*Correspondence: nirb@tauex.tau.ac.il
DOI 10.1016/j.str.2008.10.017

SUMMARY

Structural genomics initiatives provide ample structures of “hypothetical proteins” (i.e., proteins of unknown function) at an ever increasing rate. However, without function annotation, this structural goldmine is of little use to biologists who are interested in particular molecular systems. To this end, we used (an improved version of) the PatchFinder algorithm for the detection of functional regions on the protein surface, which could mediate its interactions with, e.g., substrates, ligands, and other proteins. Examination, using a data set of annotated proteins, showed that PatchFinder outperforms similar methods. We collected 757 structures of hypothetical proteins and their predicted functional regions in the N-Func database. Inspection of several of these regions demonstrated that they are useful for function prediction. For example, we suggested an interprotein interface and a putative nucleotide-binding site. A web-server implementation of PatchFinder and the N-Func database are available at http://patchfinder.tau.ac.il/.

INTRODUCTION

There is a growing need for the automatic annotation of proteins of unknown function, termed “hypothetical proteins” (Lubec et al., 2005), the structures of which are known (Friedberg, 2006). The structures of many hypothetical proteins are solved in pipelines at structural- genomics centers, which usually lack the resources to engage in thorough functional characterization of each of the solved structures. Moreover, some of the proteins, which are considered to be well annotated, may have additional functions beyond their listed records (e.g., moonlighting protein functions or promiscuous enzymatic capabilities [Copley, 2003]).

Previous attempts to collect and annotate hypothetical proteins have resulted for example, in the PDB-UF database (von Grotthuss et al., 2006), the ProKnow server (Pal and Eisenberg, 2005), and the ProFunc server (Laskowski et al., 2005a). The PDB-UF database is a collection of all of the structural genomics proteins whose functions are recorded in the Protein Data Bank (PDB) file as “unknown.” Some of the enzymes in the database have been assigned Enzyme Commission (EC) numbers (http://www.chem.qmul.ac.uk/iubmb/enzyme/) on the basis of their global structural similarity to enzymes of known function. The ProKnow server has integrated a database that includes function predictions for all of the structural genomics proteins. Recently, Watson et al. (2007) established a new database by applying the ProFunc (Laskowski et al., 2005a) server for automated function annotation on the structures that emerged from the Midwest Center for Structural Genomics (MCSG). Nevertheless, as far as we know, there is still no comprehensive database of hypothetical proteins that incorporates data from external databases, thereby enabling users to determine whether functional annotations are indeed missing. In addition, all of the structures incorporated into existing databases were solved in the context of structural genomics projects, whereas structures of hypothetical proteins can result from other sources as well. In an attempt to overcome these limitations, we constructed the N-Func database, presented below.

Typically, the function(s) of a newly discovered protein may be inferred from a sequence homolog (e.g., by using BLAST) (Altschul et al., 1997), from structurally related proteins (e.g., using structural alignment tools [Wolfson et al., 2005]), or from sequence motifs (Lee et al., 2007). Alternatively, function may be inferred on the basis of properties that are associated with a specific functional class of proteins, such as, the enrichment in basic residues and the presence of specific structural motifs that characterize DNA-binding proteins (Lubec et al., 2005). When these approaches fail, focusing on the functionally important region(s) of the protein may help to characterize the protein’s functionality (Wei and Altman, 1998). Hence, identification of functional regions by various methods (Aloy et al., 2001; Innis et al., 2004; Landgraf et al., 2001; Madabushi et al., 2002; Nimrod et al., 2005; Ondrechen et al., 2001; Pazos and Sternberg, 2004; Pettit et al., 2007) could be the first step toward function annotation. The next step would be, for example, to match the predicted functional region with a known functional site in another protein, by using, e.g., SiteEngine (Shulman-Peleg et al., 2004) and other analytical tools. Some examples are provided below.

In addition to function annotation, the identification of functionally important regions in proteins is useful for mutation analysis and drug discovery. Progress in computational methods for drug design along with the activity of structural genomics centers, such as the Protein Structure Initiative (PSI), have greatly motivated the development of automated methods for that task. Sequence and structure conservation (Panchenko et al., 2004; Pugalenthi et al., 2007; Stern et al., 2007; Via et al., 2007), physicochemical characteristics (Ko et al., 2005), surface curvature (Liang et al., 1998), and other properties (Amitai et al., 2004;

Kufareva et al., 2007) are commonly used to this end. Additionally, some methods specialize in the identification of specific classes of functional regions, such as enzyme active sites (Gutteridge et al., 2003; Tong et al., 2008), DNA-binding residues (Kuznetsov et al., 2006; Tsuchiya et al., 2004), and protein-protein interfaces (Elcock and McCammon, 2001; Negi et al., 2007; Ofran and Rost, 2007).

Of the properties used for the identification of functional regions, evolutionary conservation is perhaps the most widely employed, both alone and in combination with other properties. In functionally important positions, the evolutionary pressure typically retains rather limited variability within protein families. This observation is well known and has been utilized in various methods, such as the Evolutionary Trace (ET) (Lichtarge et al., 1996; Mihalek et al., 2004) and the Rate4Site algorithm (Mayrose et al., 2004; Pupko et al., 2002), implemented in ConSurf (Goldenberg et al., 2009; Landau et al., 2005).

Functionally important regions are often visible when evolutionary data are mapped on a protein's three-dimensional (3D) structure (Landau et al., 2005; Morgan et al., 2006). Usually, it is possible to detect clusters of conserved residues, corresponding to the proteins' functional regions (Landgraf et al., 2001; Panchenko et al., 2004). Several algorithms have been developed on the basis of this property (see, e.g., Aloy et al., 2001; Dean and Golding, 2000; Innis et al., 2004; Madabushi et al., 2002). Over the past years, we have been developing the PatchFinder algorithm for the identification of functional regions on the protein's surface (Nimrod et al., 2005). Generally speaking, PatchFinder searches for the largest and most highly conserved clusters of surface residues in the protein, which presumably represent the catalytic and/or binding sites. Here, we present new methodological improvements introduced into PatchFinder. The new version of PatchFinder is available as a webserver (http://patchfinder.tau.ac.il). We showed that the new version of PatchFinder outperforms its previous version and related methods by using a test set of 110 protein structures with residues annotated as functional sites (del Sol Mesa et al., 2003). In order to detect the functional regions in hypothetical proteins of known structure by using the PatchFinder algorithm, we established the N-Func database presented here. N-Func is a collection of 757 proteins of known 3D structure but unknown function whose close homologs also lack function annotation. The accompanying website provides easy access to the proteins' functional sites as predicted by PatchFinder.

RESULTS

We have developed PatchFinder, a Maximum Likelihood (ML) algorithm for the identification of functional regions on a protein's surface (Nimrod et al., 2005). The algorithm uses the PDB file of the protein's 3D structure and a Multiple Sequence Alignment (MSA) of the protein and its sequence homologs. It comprises the following three steps: (1) assignment of an evolutionary conservation score to each amino acid position based on its evolutionary rate among the homologous proteins (Mayrose et al., 2004); (2) extraction of the protein's solvent-accessible residues, with the aim of excluding residues that are conserved due to structural constraints and are usually buried in the protein core; and (3) identification of the most significant cluster of conserved residues on the protein's surface, based on the hypothesis that this ML patch is the protein's main functional region. Once the ML patch is found, the search procedure is continued for nonoverlapping, secondary functional regions, which present a weaker conservation signal.

In the new version of PatchFinder, described in detail in Experimental Procedures, the Delaunay triangulation (Barber et al., 1996; de Berg et al., 2000) was used to describe the neighborhood and accessibility to solvent of each residue. These properties were previously calculated via a simple distance measure and the residue's accessible surface area, respectively. In addition, here we computed the evolutionary conservation by using the Bayesian version of Rate4Site (Mayrose et al., 2004). This version is evidently superior to the ML version (Pupko et al., 2002), especially when the number of available homologous sequences is small.

Detection of Functional Regions: Performance Analysis

We examined PatchFinder's performance in comparison with the previous version and other methods by using a test set of 110 protein structures with residues annotated as functional sites (del Sol Mesa et al., 2003). The test set is referred to as dSM, from the name of the first author of reference (del Sol Mesa et al., 2003).

Each protein in the dSM data set included documentation of functionally important amino acids, referred to as "SITE" residues within the PDB file. PatchFinder identified at least one of the SITE residues in 95 out of 110 proteins in the test set. In 66 of the cases, at least half of the SITE residues were found. We also analyzed the predictions of PatchFinder by using a D-value measure, which is based on the distance between the predicted patch and the documented functional site (see Supplemental Data available online). For 77 of the 110 proteins in the data set, the ML patches of the PatchFinder algorithm were assigned D-values below 0.103, which we considered as successful prediction (see Supplemental Data).

First, we conducted a comparison that showed that PatchFinder is superior to several sequence-based methods (del Sol Mesa et al., 2003). The data are presented in Supplemental Data.

Next, we compared the predictions of PatchFinder with its ancestral version (Nimrod et al., 2005) and three additional applications for the prediction of functionally important sites in proteins of known 3D structure, namely, siteFiNDER|3D (Innis, 2007), the ET Viewer (Morgan et al., 2006), and HotPatch (Pettit et al., 2007). In the following paragraphs, we briefly describe each of these methods.

siteFiNDER|3D (Innis, 2007) is based on the conserved functional group (CFG) analysis that was developed by Innis et al. (2004). Briefly, the CFG analysis identifies, within the query protein, positions with evidence of evolutionary pressure to retain specific functional/chemical groups. Potential functional sites are then identified as spherical regions enriched with these predicted functional positions.

The ET Viewer is a server for an automated evolutionary analysis of proteins of known 3D structure (Morgan et al., 2006). As part of this analysis, the evolutionary importance of each residue is evaluated by using a real-value variant of the ET method (Mihalek et al., 2004). The ET Viewer provides cluster analysis of residues at various cutoffs of evolutionary importance rank. These clusters are of potential functional or structural significance (Madabushi

Table 1. PatchFinder Performance in Comparison with Other Methods

	The Fraction of SITE Residues Detected in the Patch ≥ 0.5[a]	At Least One SITE Residue Detected in the Patch[b]	Average Patch Size ± Standard Deviation[c]
PatchFinder 2008	0.61 (66)	0.87 (95)	19.43 ± 21.29
PatchFinder 2005	0.64 (70)	0.88 (96)	29.56 ± 33.58
siteFiNDER|3D	0.48 (52)	0.87 (95)	27.05 ± 24.01
ET Viewer	0.53 (58)	0.87 (95)	20.56 ± 21.67
HotPatch (best patch)	0.17 (19)	0.47 (51)	6.92 ± 7.2
HotPatch (all patches)	0.24 (26)	0.69 (75)	12.9 ± 10.5

[a] The fraction of cases for which at least half of the SITE residues were in the predicted functional patch/cluster (i.e., $[\text{SITE} \cap \text{patch}]/\text{site} \geq 0.5$). The number of cases in the dSM data set is provided in parentheses.

[b] The fraction of proteins for which there is some overlap between the ML patch and the SITE residues (i.e., $[\text{SITE} \cap \text{patch}] > 0$).

[c] The average number of residues predicted as functionally important. The results were measured for siteFiNDER|3D, HotPatch, the ET Viewer, the original version of PatchFinder (namely, PatchFinder 2005 [Nimrod et al., 2005]), and the new version of PatchFinder (PatchFinder 2008). Note that PDB ID 1nox includes only a single non-amino acid documentation of SITE (namely, the flavin mononucleotide molecule). It was therefore excluded from the calculations in the two middle columns.

et al., 2002; Mihalek et al., 2006). Both the ET Viewer and siteFiNDER|3D receive as input the 3D structure of the query protein and an MSA of homologous proteins.

The HotPatch algorithm searches for surface patches of exceptional physicochemical properties in proteins with known 3D structure (Pettit et al., 2007). HotPatch uses the protein's 3D structure alone as input. Hence, it is suitable for every protein with known 3D structure regardless of the availability of sequence homologs. The algorithm has several variants specialized in different functional categories (e.g., protesases, kinases, and transferases). When the functional class of the protein is known, using the corresponding specialized variant often yields better predictions (Pettit et al., 2007). Additionally, HotPatch can be used to analyze oligomers within the context of their oligomerization state (when available) and to utilize this information to find functional regions that are composed of several chains. Here, we used HotPatch in the variant of a generic functional site, and we analyzed the proteins as monomers in order to compare the algorithms with the same input.

As a comparison between the original version of PatchFinder, siteFiNDER|3D, the ET Viewer, HotPatch, and the new version of PatchFinder, we measured the fraction of SITE residues that were detected by each method and the total number of residues that were predicted to be functionally important. (Special considerations that were used in this analysis are detailed in Supplemental Data.) Table 1 summarizes the comparison between PatchFinder and the other methods. The analysis showed that whereas PatchFinder found at least half of the SITE residues in 61% of the proteins, siteFiNDER|3D and the ET Viewer succeeded only in 48% and 53% of the cases, respectively. The patches predicted by HotPatch are, on average, considerably smaller then those predicted by PatchFinder and the other methods. However, these patches comprised at least half of the SITE residues in up to 24% of the cases. According to the D-value measure, in 37 of the 110 proteins in the data set, the best patch found by HotPatch was a successful prediction, compared to 77 successful predictions of PatchFinder.

A comparison between the original and new versions of PatchFinder showed that the patches of the latter are 34% smaller, on average, than the former. The number of cases in which at least half of the SITE residues are in the patch decreased, on the other hand, by only 6% in the new PatchFinder. This and the D-value analysis, provided in Supplemental Data, show that the new version is significantly superior to the original version of PatchFinder, and that it provides patches that are considerably more focused on the functional region.

A Z-Score Measure of Confidence

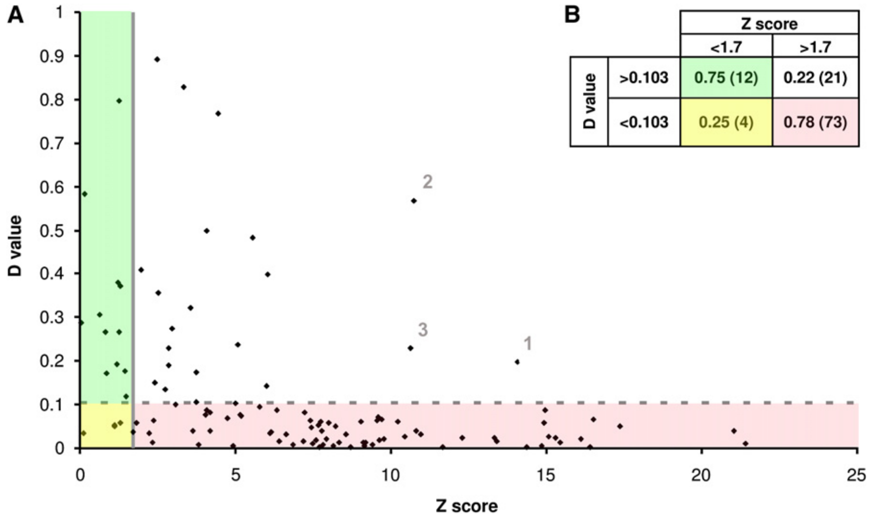
The new version of PatchFinder assigns a Z-score to each patch. The Z-score of a patch of size x with an average conservation score of y corresponds to the probability of choosing, at random, a patch of size x or larger with a conservation score equal to, or greater than, y. In Figure 1A, the D-values of the ML patches are plotted against their assigned Z-scores. The graph clearly shows that the Z-score tends to increase as the D-value decreases (Spearman correlation, $r = -0.55$; $p < 0.0001$). A similar trend was reported with the ET method (Mihalek et al., 2003).

We examined the outliers of the graph, in particular the three patches tagged "1" to "3" in Figure 1 (see Supplemental Data). In these cases, both the D-value and the Z-scores are high, which would indicate bad predictions that were assigned high confidence. The analysis showed that the high D-value may occur due to incomplete documentation of functionally important residues in the PDB file. Hence, the prediction of PatchFinder on these three cases is more successful than indicated by the D-value.

Based on these results, we used the Z-score, which is computed without knowledge of the true functional region, in order to assign a level of confidence to each individual prediction. High confidence was assigned to patches with Z-scores above 1.7 (Figure 1). This threshold was chosen with the objective of reaching 85% coverage. With this cutoff, 78% of the predictions were correct according to the gold standard. Of the proteins whose Z-scores were below 1.7, only 25% of the predictions were correct.

N-Func

We assembled the N-Func database of 757 proteins of known 3D structure but unknown function (and whose close homologs do not include function annotation). As anticipated, most of the structures (91%) had been solved in the context of worldwide structural genomics initiatives. Of the structures in N-Func,



		Z score	
		<1.7	>1.7
D value	>0.103	0.75 (12)	0.22 (21)
	<0.103	0.25 (4)	0.78 (73)

Figure 1. Evaluation of the Performance of PatchFinder by Using the dSM Test Set

(A) Scatter plot of the correlation between the Z-scores and D-values that were assigned to ML patches in the dSM test set. Predictions with a D-value smaller than 0.103 (gray, dashed line) were considered to be correct. Predictions that were assigned a Z-score greater than 1.7 (gray line) were labeled as "high confidence." The numbers 1, 2, and 3 mark the three outliers discussed in Supplemental Data.

(B) Summary of the data in (A). Values are the fractions of patches in each Z-score category; colors indicate corresponding categories. The numbers of patches within each category are recorded in parentheses. The figure shows that the accuracy of the prediction increases with increasing Z-score values.

85% had been solved by X-ray crystallography, and the rest had been solved by NMR. This is close to the ratio between structures solved by X-ray and by NMR in the entire PDB database. In addition, almost two-thirds of the structures are composed of between 150 and 249 residues, corresponding to one or two structural domains of average size (Shen et al., 2005). Of the proteins in N-Func, 66% are derived from bacteria, 17% from archaea, 16% from eukaryotes, and 1% from viruses and phages (see Figure S6).

Analysis of these proteins by PatchFinder showed that the average size of the ML patches in the database was 17 (±12) residues. Of the ML patches, 90% (681) are high-confidence predictions (Z-score > 1.7). These high Z-scores indicate that although the proteins in N-Func lack annotations, their conservation profiles make it possible to predict their functional regions. This is a first step toward function annotation of these proteins.

The N-Func database, which will be updated periodically, is available as a website (http://patchfinder.tau.ac.il/N-Func/). Each protein is allocated a "result" page, which includes an interactive 3D visualization of the significant patches (using FirstGlance in Jmol), the calculated conservation scores of each amino acid position in the protein, and the input MSA.

Function is known to be transferable between sequence homologs (Rost et al., 2003), but the measure of similarity for reliable transfer of function between proteins is still a matter of debate (Devos and Valencia, 2000; Rost, 2002; Thornton, 2001). By utilizing the N-Func database, the user can retrieve proteins according to a preset sequence-identity cutoff of between 30% and 95% to a homologous protein with known function. As an example, by choosing a sequence-identity threshold of 45%, the user can view a list of all N-Func entries for which the UniProt database (Bairoch et al., 2005) contains no functional annotation for the protein, as well as for any homolog with a sequence identity of 45% or more. Currently, that list includes 594 of the 757 proteins in N-Func.

Function Annotation: Three Cases

Manual inspection of some of the patches in the N-Func database convinced us that they are functionally important indeed. Three examples are presented below, and a fourth one is available in Supplemental Data.

Q9I5E5_PSEAE: PatchFinder's Functional Site Is Supported by Additional Evidence

PDB ID 2h9f refers to a hypothetical protein of 391 residues from *Pseudomonas aeruginosa*. PatchFinder analysis revealed an ML patch of 24 residues with a Z-score of 11.1 (Figure 2A, red) and a secondary patch of 19 residues with a Z-score of 2.14 (green). These two patches delineate the functional site in the largest cavity of the protein. Mapping of the electrostatic potential (Baker et al., 2001) on the protein's surface shows that the cavity's potential is conspicuously positive (Figure 2B). The presence of such highly conserved patches of amino acids in (and near) a highly charged cavity strongly supports the hypothesis that the two patches are located in the protein's functional site. It further suggests that the protein binds to a large negatively charged molecule. A more detailed analysis is provided in the Supplemental Data.

Q8E989_SHEON: Biological Interfaces and Crystal Contacts

PatchFinder analysis can also be used to examine the physiological relevance of various crystallographic interfaces between proteins. Some of the observed protein-protein interfaces might reflect crystal packing, whereas others could be genuine contacts of functional importance. Evolutionary data can be utilized to discriminate between them (Elcock and McCammon, 2001; Valdar and Thornton, 2001); real interfaces are usually more conserved, as exemplified in the case of PDB ID 1t82 from *Shewanella oneidensis*. The crystal structure shows a homotetrameric protein with two distinct protein-protein interfaces. PatchFinder analysis showed that the ML patch overlaps with one of the interfaces (Figure S5), suggesting that it is physiological. This interface is symmetrical and involves contacts between the ML patch of each of the two subunits. In contrast, the residues that comprise the second crystallographic interface are highly variable, suggesting that it is nonphysiological.

The PQS server (Henrick and Thornton, 1998) differentiates between contacts that are biologically relevant and those reflecting crystal packing. The prediction of the PQS server is based on several parameters, including the size of the interface and the

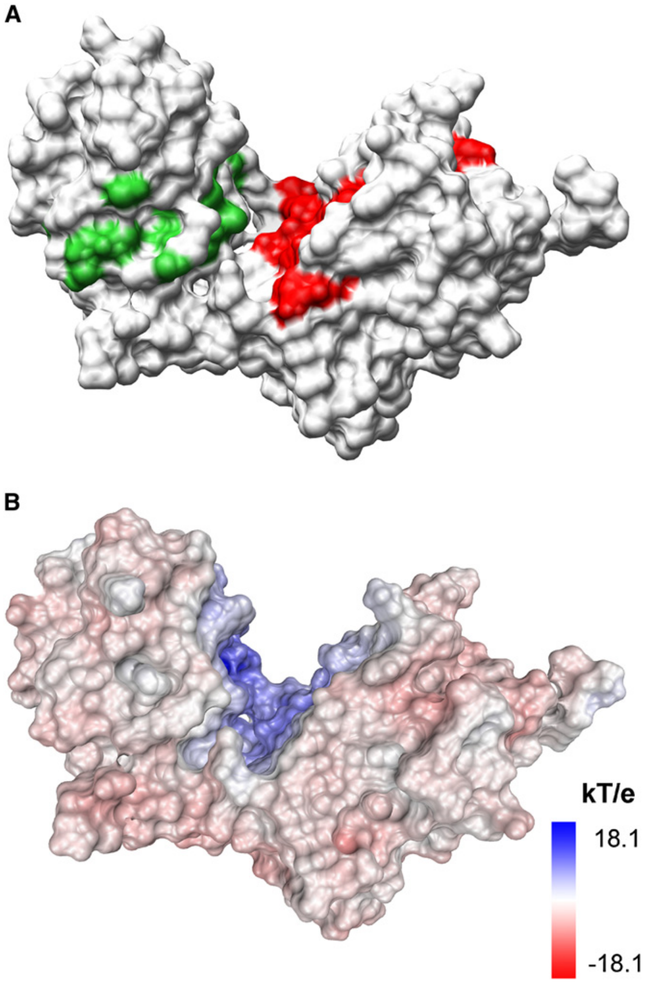


Figure 2. The Conserved and Charged Cavity in Hypothetical Protein Q9I5E5_PSEAE

(A) The ML patch and the secondary patch are colored green and red, respectively. The rest of the protein is gray.

(B) Projection of the electrostatic potential on the protein surface, using the color bar on the right. The figure was produced by using PMV (Sanner, 1999). The evolutionarily conserved patches are evidently located in the largest cavity of the protein, which is positively charged. Overall, the analysis indicates that PatchFinder detected a region of association with a negatively charged molecule.

estimated difference in solvation energy between the protein's dimeric and monomeric forms. Evolutionary conservation is not taken into consideration; therefore, PQS calculations are complementary to PatchFinder. The prediction of the PQS server supports our supposition that 1t82 dimerizes through the interface region delineated by the ML patch.

Further support for the biological relevance of the interface comes from the cons-PPISP server (Chen and Zhou, 2005). This server is based on a neural network method that uses as input the position-specific sequence profile (Altschul et al., 1997) and the solvent accessibility of the examined positions. cons-PPISP predicted a surface cluster of 20 residues as a site of protein-protein interaction. PatchFinder found 19, of which 15 overlap with the cons-PPISP cluster.

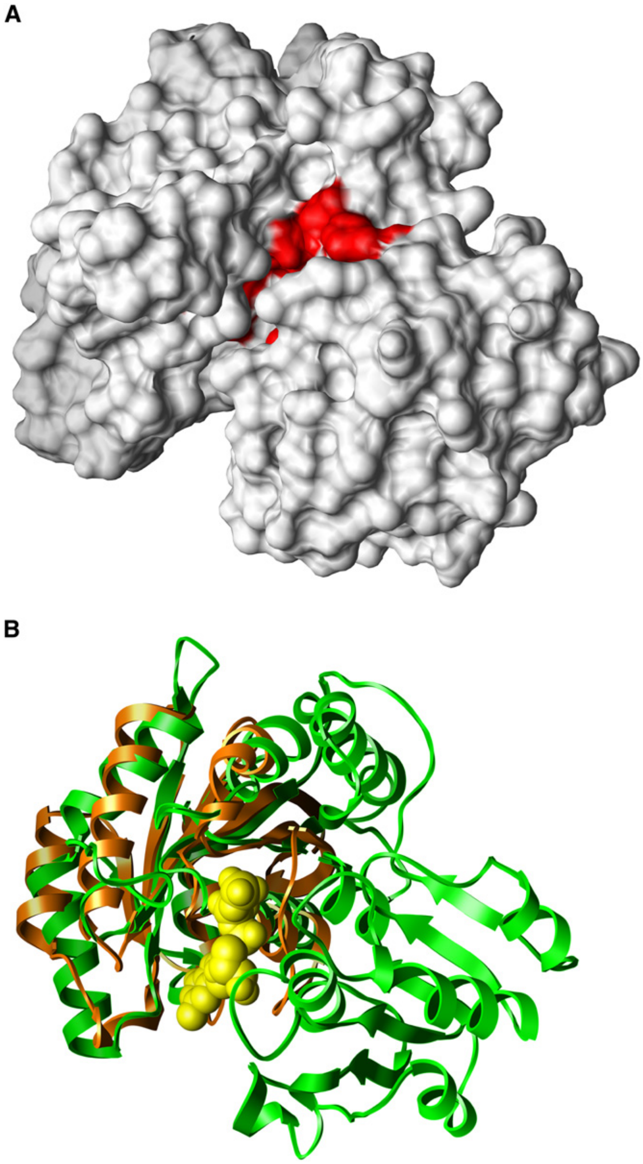


Figure 3. Hypothetical Protein Q3M7B8_ANAVT from *Anabaena variabilis*

(A) Surface representation of the protein. The patch is colored red, and the rest of the protein is gray.

(B) Superimposition (Laskowski et al., 2005b) of the structure of Q3M7B8_ANAVT (green ribbons) and the structure of human Rab5a (orange ribbons [Zhu et al., 2003]) with a GNP molecule (yellow, spacefilled representation). Clashes between the GNP molecule and Q3M7B8_ANAVT appear to be only minor, suggesting that Q3M7B8_ANAVT might bind nucleotides.

Q3M7B8_ANAVT: A Potential GTPase

Another interesting protein in the database is Q3M7B8_ANAVT from *Anabaena variabilis*, as represented by PDB ID 2obn. The protein forms homodimeric structures that are predicted to be biologically relevant according to the PQS server (Henrick and Thornton, 1998). PatchFinder analysis revealed an ML patch of 15 residues with a Z-score of 12.8. This patch resides within the largest cavity of the protein (Figure 3A), a location that often delineates a possible ligand-binding site (Liang et al., 1998). We used the ProFunc server (Laskowski et al., 2005a) to examine the protein's functionality. ProFunc revealed marked local structural similarity (Laskowski et al., 2005b) between the query protein and the ligand-binding sites of several GTP-binding proteins, including human Rab5a (Zhu et al., 2003). The local structural similarity of Rab5a's GTP-binding site to a subregion of 2obn was used to superimpose the entire structures. The superimposition revealed a strong resemblance between the global fold of the human Rab5a and one wing of the Q3M7B8_ANAVT structure (Figure 3B). In addition, the GTP analog, GNP, could be seen to reside within the cavity of Q3M7B8_ANAVT with only minor clashes.

Further analysis, presented in Supplemental Data, suggests that Q3M7B8_ANAVT is likely to bind GTP or another free nucleotide. Experimental investigation of Q3M7B8_ANAVT is needed in order to examine the predictions, elucidate the mechanism of nucleotide hydrolysis in detail, and study the roles of the residues from both of the relevant domains.

DISCUSSION

Here, we presented an improved version of PatchFinder, based on the incorporation of the Delaunay triangulation and the Bayesian version of the Rate4Site algorithm. The results demonstrated significant improvements over the old version, attributable, for the most part, to a substantial increase in precision at the cost of only a minor decrease in recall (see Supplemental Data). The new version of PatchFinder also provides an estimate of the level of confidence of the prediction.

Based on the SITE annotation in the PDB (and the D-values < 0.103 cutoff), we successfully detected 77 of the 110 functional sites in the dSM test set, corresponding to a detection rate of 70%. However, as described above, examination of the other cases showed that, in many instances, the ML patch corresponded to a functional region that was not annotated in the PDB. In principle, it is not implausible that many of the predicted ML patches that were calculated on the basis of a sufficient number of homologous proteins represent functional regions that have yet to be discovered. In this respect, some of the apparently false predictions can still be considered as functionally important regions, and the actual detection rate of PatchFinder is likely to be considerably higher than 70%. A list of the amino acids comprising each of the ML patches and those predicted by the related methods is provided at the accompanying website. We are hopeful that experimentalists will test the predictions.

Comparing PatchFinder with Other Methods

The comparison between PatchFinder and similar methods has a few marked limitations. One limitation is based on the way in which the "functionally important regions" of a protein are defined. An important assumption in our development of PatchFinder was that the functionally important regions comprise only amino acids that mediate interactions with other biomolecules. Based on this definition, PatchFinder overlooks residues that, though highly conserved, are completely buried in the protein core. An alternative approach is aimed at the "identification of evolutionarily important residues" (Lichtarge et al., 2003), regardless of the extent to which they are buried. Thus, buried residues that are highly conserved because of their role in stabilization of the 3D structure of the protein may be detected by methods that are based on that broader definition. Yet another approach looks for the catalytic residues (Petrova and Wu, 2006), which constitute a subgroup of the residues included in our definition. Methods based on such different definitions of the proteins' functional regions naturally produce dissimilar results, and comparisons between them may be misleading.

Moreover, a comprehensive comparison requires a data set of proteins that have undergone thorough mutation analysis, enabling a clear distinction between functional regions and the rest of the protein (Mihalek et al., 2004). Such data, however, are available in only a few cases, such as the *E. coli* lactose repressor (Markiewicz et al., 1994).

In spite of these limitations, we compared PatchFinder with related methods. Our analysis showed that PatchFinder performs better than the other methods on the dSM test set. We demonstrated that PatchFinder finds a considerable part of the SITE in more cases than siteFiNDER|3D and the ET Viewer. Moreover, the average size of the functional region predicted by PatchFinder was 28% smaller than that of siteFiNDER|3D. This is indicative of the superior precision of PatchFinder.

Unlike PatchFinder and the other methods compared here, HotPatch (Pettit et al., 2007) is not based on the evolutionary conservation of the residues of the query proteins. Instead, it uses various physicochemical properties, such as concavity and hydrophobicity. HotPatch found patches that were considerably smaller, on average, than those identified by PatchFinder. On the other hand, HotPatch did not find SITE residues in nearly one-third of the cases, which is indicative of inferior sensitivity in comparison with PatchFinder, at least in the dSM data set. In this respect, it should be mentioned that HotPatch was developed to maximize specificity, which is difficult to measure in this study because of the incompleteness of the documentation of SITE residues in the dSM data set. In our analysis, the specificity is estimated, indirectly, based on the size of the patch/cluster.

The approach used by HotPatch is complementary to that of PatchFinder (and the ET Viewer and the siteFiNDER|3D methods). Therefore, regions identified by both methods as functionally important are more likely to be true predictions. Furthermore, an algorithm that will combine the properties examined by PatchFinder and HotPatch may improve their performance.

Regardless of performance, HotPatch and PatchFinder are unique in that they are fully automatic and report a single (or several) patch(es), which presumably corresponds to the functional region(s) of the protein. The user is not required to make any decisions along the way. Having said that, it is important to note that, in the HotPatch and PatchFinder webservers, the user may tailor the computational protocol to a specific need. For example, the user can use a protocol that was found to be most suitable for the detection of specific catalytic sites in HotPatch or choose the number and type of homologs that are used for the PatchFinder analysis. The other methods provide a list of putative clusters without ranking them by the likelihood to be functional. The ET Viewer enables the user to tune the evolutionary importance cutoff for the clustering in order to suit the analysis to the specific examined case. In our analysis, we virtually enforce an automated cluster selection for the ET viewer (see Supplemental Data). This procedure did not necessarily yield the best selection of clusters for each protein; manual inspection may have improved the performance of the ET Viewer.

Limitations and Implications of PatchFinder and N-Func

The use of evolutionary conservation might be impractical in some cases. For example, when only a few sequence homologs for the query protein are available, it might be difficult to track the evolutionary process within the family. In such cases, the evolutionary signal is weak, and it might be problematic to differentiate between amino acid positions that evolve slowly because of negative selection and hence comprise the functional region (i.e., signal) and those that appear to be conserved because of the short evolutionary time (i.e., noise).

It is also important to note that the functional region might not always be evolutionarily conserved. In extreme cases, such as the peptide-recognition regions of antibodies and MHC molecules, the functional region is highly variable (Reche and Reinherz, 2003).

The novelty of PatchFinder, compared to similar methods, comes from two main aspects. First, it distinguishes between positions that are functionally important and those that are structurally important. Second, the identification and delineation of the patches are conducted within the framework of the ML approach, which is statistically robust.

PatchFinder was implemented here as a fully automated webserver, which is easy to use and enables the user to provide MSAs (in any of the common formats) and predict functional sites in proteins of their interest. This might prove useful for the design and interpretation of mutagenesis studies, the rational design of protein and drug, and the interpretation of genetic and clinical data.

Annotation of Protein Function

The analysis of ML patches might also be useful for the prediction of a protein's function. As an example, a preliminary analysis in our laboratory showed convincingly that the ML patches of DNA-binding proteins differ significantly in amino acid composition and electrostatic potentials from those of other proteins (G. Nimrod, A. Szilágyi, C. Leslie, and N. Ben-Tal, unpublished data). It should therefore be possible to identify DNA-binding proteins on the basis of the properties of their ML patches.

N-Func is a collection of hypothetical proteins that were automatically analyzed by using the (improved) PatchFinder algorithm. This database currently provides the ML patches of 757 PDB entries, thus serving as an initial step toward function annotation of these proteins. The ML patches of 90% of these proteins were assigned a high level of confidence. This, together with the detailed examples presented above, indicates that N-Func provides valuable data that may ultimately be used to suggest the functions of these proteins. This will most probably be done by integrating PatchFinder predictions with data obtained by the use of other computational tools. It should be possible, for example, to characterize these proteins by looking for local similarities between the predicted functional regions and known sites in other proteins (Laskowski et al., 2005b; Shulman-Peleg et al., 2004). The ML patches are also potentially useful in designing experiments for the deduction of functionality.

EXPERIMENTAL PROCEDURES

Collection of Proteins for N-Func

Our objective in constructing the N-Func database was to gather proteins with available 3D structures but with no functional annotations. We based our search on the PDB and UniProt (Bairoch et al., 2005) documentation for each protein. Using a text-based search in the RCSB website (http://www.pdb.org/), we first listed all PDB entries that contained the terms "hypothetical" or "unknown function." This yielded an initial collection of 2245 entries. To avoid duplicates and filter the structures, we used the PISCES server (Wang and Dunbrack, 2003) to cull the initial collection by defining the following three structural parameters as inclusion criteria: resolution of, at most, 3 Å; maximum sequence identity of 99% between all chains; and minimum chain length of 100 amino acids (this threshold was chosen in order to avoid incomplete structural domains).

Culling reduced the initial collection to 1599 nonidentical protein chains. We also removed 43 entries of hetero-oligomeric structures since, by definition, these cases present annotations that relate to interprotein interactions. In addition, we removed 71 entries with fewer than four homologous sequences in the Homology-derived Secondary Structures of Proteins (HSSP) database (Schneider and Sander, 1996), because Rate4Site often fails to detect the evolutionary signal in such cases (Mayrose et al., 2004).

For the final filtering step, we checked the presence or absence of functional annotations provided by the UniProt database (Bairoch et al., 2005). The examined fields in the UniProt entry of each protein were "function," "catalytic activity," "GO" (Ashburner et al., 2000), and "EC." Only proteins for which all of these fields were missing were considered in N-Func. Furthermore, if any close sequence homolog (>95% sequence identity over >80% of the protein length) of a particular protein had functional annotation, that protein was also excluded from the database since function can safely be inferred from the homologs in such cases.

Multiple Sequence Alignments

Rate4Site (Mayrose et al., 2004) computes the conservation score for each amino acid position in the protein based on the MSA of the query protein family. The input MSA was extracted from the HSSP database (Schneider and Sander, 1996) by using MVIEW (Brown et al., 1998). Parameterization and improvements in PatchFinder were first introduced in 2005, with the dSM test set and their matching MSAs from the HSSP compilation of that year. The MSAs are provided at the PatchFinder website. The predictions in N-Func are based on a more recent release of HSSP (in 2007).

Solvent Accessibility and Connectivity

The solvent accessibility of each residue and the identities of its surrounding residues were determined by using the Delaunay triangulation (Barber et al., 1996; de Berg et al., 2000). The center of each heavy atom in the query protein was considered as a vertex. The 3D Delaunay triangulation created nonoverlapping tetrahedral shapes, where each atom center was a vertex of at least one tetrahedron. The shapes assembled into a convex hull enclosing the protein.

We considered a vertex to be on the protein surface if a face that it belonged to appeared in exactly one tetrahedron. Some of the tetrahedral shapes, being within surface cavities (Liang et al., 1998), typically had long edges. Therefore, tetrahedra with edges longer than a certain cutoff distance (see below) were iteratively removed, exposing the "floor" of the cavity. A residue was considered to be exposed if at least one of its atoms was a surface vertex.

Two residues were considered to be adjacent if they had surface atoms that shared the same tetrahedron and the distance between these atoms was smaller to a preset cutoff distance. The distance chosen was the sum of a probe with a diameter of 2.8 Å, corresponding to a water molecule, and the van der Waals radii of the atoms (Chothia, 1976).

There are various measures by which to determine the solvent accessibility of each residue/atom in a protein structure quantitatively (for example, the accessible surface area [Lee and Richards, 1971] that was used in the first version of PatchFinder). Here, we used the Delaunay triangulation for consistency with the description of the neighborhood of each atom. Furthermore, the current procedure determines whether an atom is exposed or buried within the protein in a binary manner and does not require the user to choose a solvent-accessibility cutoff.

Generating Structural Figures

All molecular graphical pictures were produced by using UCSF Chimera (Pettersen et al., 2004), except for Figure 2B, which was produced with the

Python Molecular Viewing environment PMV (Sanner, 1999), and Figure S5, which was generated with PyMol (DeLano, 2002).

SUPPLEMENTAL DATA

Supplemental Data include Supplemental Experimental Procedures, Supplemental References, and three figures and can be found with this article online at http://www.cell.com/structure/supplemental/S0969-2126(08)00423-1.

ACKNOWLEDGMENTS

We thank Frank Pettit, Axel Innis, Angela Wilkins, Ivana Mihalek, and Olivier Lichtarge for applying their algorithms to the data set and for helpful discussions. We would like to acknowledge Leonid Brodsky, Sarel Fleishman, Yanay Ofran, Gilad Wainreb, Lei Xie, and Tsaffrir Zor for helpful discussions. This work was supported by the European Commission 6th Framework Programme (FP6) Integrated Project EuroHear, LSHG-CT-20054-512063.

Received: April 16, 2008
Revised: October 16, 2008
Accepted: October 19, 2008
Published: December 9, 2008

REFERENCES

Aloy, P., Querol, E., Aviles, F.X., and Sternberg, M.J. (2001). Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J. Mol. Biol. *311*, 395–408.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I., and Pietrokovski, S. (2004). Network analysis of protein structures identifies functional residues. J. Mol. Biol. *344*, 1135–1146.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29.

Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. (2005). The Universal Protein Resource (UniProt). Nucleic Acids Res. *33*, D154–D159.

Baker, N.A., Sept, D., Joseph, S., Holst, M.J., and McCammon, J.A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. Proc. Natl. Acad. Sci. USA *98*, 10037–10041.

Barber, C.B., David, P.D., and Hannu, H. (1996). The quickhull algorithm for convex hulls. ACM Trans. Math. Softw. *22*, 469–483.

Brown, N.P., Leroy, C., and Sander, C. (1998). MView: a web-compatible database search or multiple alignment viewer. Bioinformatics *14*, 380–381.

Chen, H., and Zhou, H.X. (2005). Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. Proteins *61*, 21–35.

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. J. Mol. Biol. *105*, 1–12.

Copley, S.D. (2003). Enzymes with extra talents: moonlighting functions and catalytic promiscuity. Curr. Opin. Chem. Biol. *7*, 265–272.

de Berg, M., van Kreveld, M., Overmars, M., and Schwarzkopf, O. (2000). Delaunay Triangulations-Height Interpolation. In Computational Geometry: Algorithms and Applications. M. de Berg, M. van Krefeld, M. Overmars, and O. Schwarzkopf, eds. (New York: Springer), pp. 183–210.

Dean, A.M., and Golding, G.B. (2000). Enzyme evolution explained (sort of). Pac. Symp. Biocomput. 6–17.

del Sol Mesa, A., Pazos, F., and Valencia, A. (2003). Automatic methods for predicting functionally important residues. J. Mol. Biol. *326*, 1289–1302.

DeLano, W.L. (2002). The PyMOL Molecular Graphics Program (http://www.pymol.org).

Devos, D., and Valencia, A. (2000). Practical limits of function prediction. Proteins *41*, 98–107.

Elcock, A.H., and McCammon, J.A. (2001). Identification of protein oligomerization states by analysis of interface conservation. Proc. Natl. Acad. Sci. USA *98*, 2990–2994.

Friedberg, I. (2006). Automated protein function prediction–the genomic challenge. Brief. Bioinform. *7*, 225–242.

Goldenberg, O., Erez, E., Nimrod, G., and Ben-Tal, N. (2009). The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. Nucleic Acids Res., in press, 10.1093/nar/gkn822.

Gutteridge, A., Bartlett, G.J., and Thornton, J.M. (2003). Using a neural network and spatial clustering to predict the location of active sites in enzymes. J. Mol. Biol. *330*, 719–734.

Henrick, K., and Thornton, J.M. (1998). PQS: a protein quaternary structure file server. Trends Biochem. Sci. *23*, 358–361.

Innis, C.A. (2007). siteFiNDER|3D: a web-based tool for predicting the location of functional sites in proteins. Nucleic Acids Res. *35*, W489–W494.

Innis, C.A., Anand, A.P., and Sowdhamini, R. (2004). Prediction of functional sites in proteins using conserved functional group analysis. J. Mol. Biol. *337*, 1053–1068.

Ko, J., Murga, L.F., Wei, Y., and Ondrechen, M.J. (2005). Prediction of active sites for protein structures from computed chemical properties. Bioinformatics *21*, i258–i265.

Kufareva, I., Budagyan, L., Raush, E., Totrov, M., and Abagyan, R. (2007). PIER: protein interface recognition for structural proteomics. Proteins *67*, 400–417.

Kuznetsov, I.B., Gou, Z., Li, R., and Hwang, S. (2006). Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. Proteins *64*, 19–27.

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., and Ben-Tal, N. (2005). ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res. *33*, W299–W302.

Landgraf, R., Xenarios, I., and Eisenberg, D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J. Mol. Biol. *307*, 1487–1502.

Laskowski, R.A., Watson, J.D., and Thornton, J.M. (2005a). ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res. *33*, W89–W93.

Laskowski, R.A., Watson, J.D., and Thornton, J.M. (2005b). Protein function prediction using local 3D templates. J. Mol. Biol. *351*, 614–626.

Lee, B., and Richards, F.M. (1971). The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. *55*, 379–400.

Lee, D., Redfern, O., and Orengo, C. (2007). Predicting protein function from sequence and structure. Nat. Rev. Mol. Cell Biol. *8*, 995–1005.

Liang, J., Edelsbrunner, H., and Woodward, C. (1998). Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci. *7*, 1884–1897.

Lichtarge, O., Bourne, H.R., and Cohen, F.E. (1996). An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. *257*, 342–358.

Lichtarge, O., Yao, H., Kristensen, D.M., Madabushi, S., and Mihalek, I. (2003). Accurate and scalable identification of functional sites by evolutionary tracing. J. Struct. Funct. Genomics *4*, 159–166.

Lubec, G., Afjehi-Sadat, L., Yang, J.W., and John, J.P. (2005). Searching for hypothetical proteins: theory and practice based upon original data and literature. Prog. Neurobiol. *77*, 90–127.

Madabushi, S., Yao, H., Marsh, M., Kristensen, D.M., Philippi, A., Sowa, M.E., and Lichtarge, O. (2002). Structural clusters of evolutionary trace residues are statistically significant and common in proteins. J. Mol. Biol. *316*, 139–154.

Markiewicz, P., Kleina, L.G., Cruz, C., Ehret, S., and Miller, J.H. (1994). Genetic studies of the lac repressor. XIV. Analysis of 4000 altered *Escherichia coli* lac repressors reveals essential and non-essential residues, as well as “spacers” which do not require a specific sequence. J. Mol. Biol. *240*, 421–433.

Mayrose, I., Graur, D., Ben-Tal, N., and Pupko, T. (2004). Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol. Biol. Evol. *21*, 1781–1791.

Mihalek, I., Res, I., Yao, H., and Lichtarge, O. (2003). Combining inference from evolution and geometric probability in protein structure evaluation. J. Mol. Biol. *331*, 263–279.

Mihalek, I., Res, I., and Lichtarge, O. (2004). A family of evolution-entropy hybrid methods for ranking protein residues by importance. J. Mol. Biol. *336*, 1265–1282.

Mihalek, I., Res, I., and Lichtarge, O. (2006). Evolutionary trace report_maker: a new type of service for comparative analysis of proteins. Bioinformatics *22*, 1656–1657.

Morgan, D.H., Kristensen, D.M., Mittelman, D., and Lichtarge, O. (2006). ET viewer: an application for predicting and visualizing functional sites in protein structures. Bioinformatics *22*, 2049–2050.

Negi, S.S., Schein, C.H., Oezguen, N., Power, T.D., and Braun, W. (2007). InterProSurf: a web server for predicting interacting sites on protein surfaces. Bioinformatics *23*, 3397–3399. Published online October 12, 2007. 10.1093/bioinformatics/btm474.

Nimrod, G., Glaser, F., Steinberg, D., Ben-Tal, N., and Pupko, T. (2005). In silico identification of functional regions in proteins. Bioinformatics *21*, i328–i337.

Ofran, Y., and Rost, B. (2007). Protein-protein interaction hotspots carved into sequences. PLoS Comput. Biol. *3*, e119.

Ondrechen, M.J., Clifton, J.G., and Ringe, D. (2001). THEMATICS: a simple computational predictor of enzyme function from structure. Proc. Natl. Acad. Sci. USA *98*, 12473–12478.

Pal, D., and Eisenberg, D. (2005). Inference of protein function from protein structure. Structure *13*, 121–130.

Panchenko, A.R., Kondrashov, F., and Bryant, S. (2004). Prediction of functional sites by analysis of sequence and structure conservation. Protein Sci. *13*, 884–892.

Pazos, F., and Sternberg, M.J. (2004). Automated prediction of protein function and detection of functional sites from structure. Proc. Natl. Acad. Sci. USA *101*, 14754–14759.

Petrova, N.V., and Wu, C.H. (2006). Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. BMC Bioinformatics *7*, 312.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612.

Pettit, F.K., Bare, E., Tsai, A., and Bowie, J.U. (2007). HotPatch: a statistical approach to finding biologically relevant features on protein surfaces. J. Mol. Biol. *369*, 863–879.

Pugalenthi, G., Suganthan, P.N., Sowdhamini, R., and Chakrabarti, S. (2007). SMotif: a server for structural motifs in proteins. Bioinformatics *23*, 637–638.

Pupko, T., Bell, R.E., Mayrose, I., Glaser, F., and Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics *18* (*suppl.1*), S71–S77.

Reche, P.A., and Reinherz, E.L. (2003). Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms. J. Mol. Biol. *331*, 623–641.

Rost, B. (2002). Enzyme function less conserved than anticipated. J. Mol. Biol. *318*, 595–608.

Rost, B., Liu, J., Nair, R., Wrzeszczynski, K.O., and Ofran, Y. (2003). Automatic prediction of protein function. Cell. Mol. Life Sci. *60*, 2637–2650.

Sanner, M.F. (1999). Python: a programming language for software integration and development. J. Mol. Graph. Model. *17*, 57–61.

Schneider, R., and Sander, C. (1996). The HSSP database of protein structure-sequence alignments. Nucleic Acids Res. *24*, 201–205.

Shen, M.-y., Davis, F.P., and Sali, A. (2005). The optimal size of a globular protein domain: a simple sphere-packing model. Chem. Phys. Lett. *405*, 224–228.

Shulman-Peleg, A., Nussinov, R., and Wolfson, H.J. (2004). Recognition of functional sites in protein structures. J. Mol. Biol. *339*, 607–633.

Stern, A., Doron-Faigenboim, A., Erez, E., Martz, E., Bacharach, E., and Pupko, T. (2007). Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Res. *35*, W506–W511.

Thornton, J.M. (2001). From genome to function. Science *292*, 2095–2097.

Tong, W., Williams, R.J., Wei, Y., Murga, L.F., Ko, J., and Ondrechen, M.J. (2008). Enhanced performance in prediction of protein active sites with THEMATICS and support vector machines. Protein Sci. *17*, 333–341.

Tsuchiya, Y., Kinoshita, K., and Nakamura, H. (2004). Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. Proteins *55*, 885–894.

Valdar, W.S., and Thornton, J.M. (2001). Conservation helps to identify biologically relevant crystal contacts. J. Mol. Biol. *313*, 399–416.

Via, A., Peluso, D., Gherardini, P.F., de Rinaldis, E., Colombo, T., Ausiello, G., and Helmer-Citterich, M. (2007). 3dLOGO: a web server for the identification, analysis and use of conserved protein substructures. Nucleic Acids Res. *35*, W416–W419.

von Grotthuss, M., Plewczynski, D., Ginalski, K., Rychlewski, L., and Shakhnovich, E.I. (2006). PDB-UF: database of predicted enzymatic functions for unannotated protein structures from structural genomics. BMC Bioinformatics *7*, 53.

Wang, G., and Dunbrack, R.L., Jr. (2003). PISCES: a protein sequence culling server. Bioinformatics *19*, 1589–1591.

Watson, J.D., Sanderson, S., Ezersky, A., Savchenko, A., Edwards, A., Orengo, C., Joachimiak, A., Laskowski, R.A., and Thornton, J.M. (2007). Towards fully automated structure-based function prediction in structural genomics: a case study. J. Mol. Biol. *367*, 1511–1522.

Wei, L., and Altman, R.B. (1998). Recognizing protein binding sites using statistical descriptions of their 3D environments. Pac. Symp. Biocomput. 497–508.

Wolfson, H.J., Shatsky, M., Schneidman-Duhovny, D., Dror, O., Shulman-Peleg, A., Ma, B., and Nussinov, R. (2005). From structure to function: methods and applications. Curr. Protein Pept. Sci. *6*, 171–183.

Zhu, G., Liu, J., Terzyan, S., Zhai, P., Li, G., and Zhang, X.C. (2003). High resolution crystal structures of human Rab5a and five mutants with substitutions in the catalytically important phosphate-binding loop. J. Biol. Chem. *278*, 2452–2460.